Jason Perlmutter - Analysis of Passenger Air Travel

30 October, 2016

Background

The Bureau of Transportation Statistics data set contains information on domestic passenger air travel. Specfically, for each flight, the data set includes the origin and destination airport, the date of travel, the duration and delay of the flight, and the airline carrier, amongst other details. This will be combined with supplemental files which provide airline name and airport geography. To make analysis of this large data set tractable, I am going to focus on a single, complete year of data (2015) and my primary interest will be looking into flight delays (which I define as arrival >15 minutes after scheduled).

As a preliminary step, I describe below the average number of flights per day, sorted by day of week, month, Airline, and Airport. The tables below show that there are fewer flights on Saturdays than other days of the week, and more flights over the summer months than winter months. The number of flights by airline and airport vary widely, and only the top 10 of each category are listed below.

These charts suggest several potential questions– For example, are there certain routes/destinations which are more common in the summer than winter? Which routes/destinations are absent from Saturday travel? Presumably, there are relationships between the number of flights at an airport and through an airline, with certain airlines favoring certain regions, which could be further elucidated.

Table 1: Flights by Day of week

Day of Week	Avg Number of Flights Per Day
Monday	16151
Tuesday	15874
Wednesday	16179
Thursday	16149
Friday	16335
Saturday	13226
Sunday	15396

Table 2: Flights by Month

Month	Avg Number of Flights Per Day
January	14706
February	14524
March	15826
April	15935
May	15758
June	16389
July	16548
August	16214
September	15373
October	15545
November	15381
December	15121

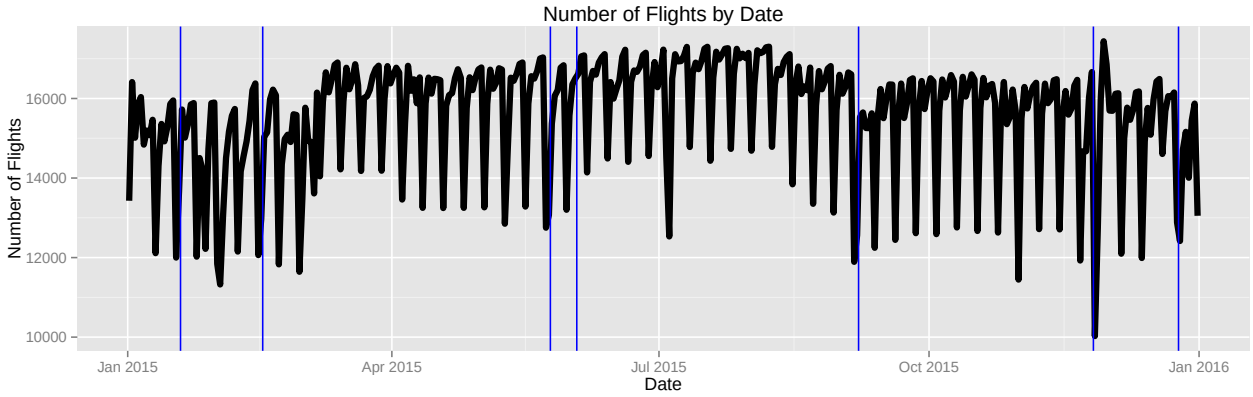
Table 3: Flights by Airline

Carrier	Avg Number of Flights Per Day
Southwest Airlines Co.	3394
Delta Air Lines Inc.	2373
American Airlines Inc.	1953
Skywest Airlines Inc.	1577
Atlantic Southeast Airlines	1511
United Air Lines Inc.	1391
American Eagle Airlines Inc.	762
JetBlue Airways	718
US Airways Inc.	532
Alaska Airlines Inc.	470

Table 4: Flights by Airport

Airport	Total Flights Per Day
William B Hartsfield-Atlanta Intl	2047
Chicago O'Hare International	1659
Dallas-Fort Worth International	1385
Denver Intl	1158
Los Angeles International	1146
San Francisco International	874
Phoenix Sky Harbor International	868
George Bush Intercontinental	856
McCarran International	793
Minneapolis-St Paul Intl	667

The plot beneath shows the number of flights for each day of 2015, with the blue lines indicating US Holidays. One hypothesis is that there are a larger number of flights around holidays, and while this appears to be true for after Thanksgiving, it does not seem to hold for other holidays.



Delay Rate

In the charts below, I describe the rate at which flights are delayed by day of week, month, airline, and airport.

It is interesting to note that there are fewer flights on Saturdays, and these flights are less often delayed. This suggests a correlation between number of flights and delays. However, this is contradicted by the rate of delays be month, where it appears that there are more delays in the summer (when there are more flights) as well as in the winter (when there are fewer flights). The full data set contains information on delay cause, and it would be interesting as a next step to investigate that as well– it seems likely that the delays in February are due to weather, while the delays in the summer or around holidays are due to high traffic.

Further, the fact that delay is more common at certain airlines, airports, or days of the week is information that would be actionable for travelers when choosing their flights.

Table 5: Delayed by Day of week

Day of Week	Delay Rate
Monday	0.19
Tuesday	0.18
Wednesday	0.18
Thursday	0.19
Friday	0.18
Saturday	0.15
Sunday	0.18

Table 6: Delayed by Month

Month	Delay Rate
January	0.20
February	0.23
March	0.19
April	0.16
May	0.18
June	0.23
July	0.20
August	0.18
September	0.12
October	0.12
November	0.15
December	0.20

Table 7: Delayed by Airline

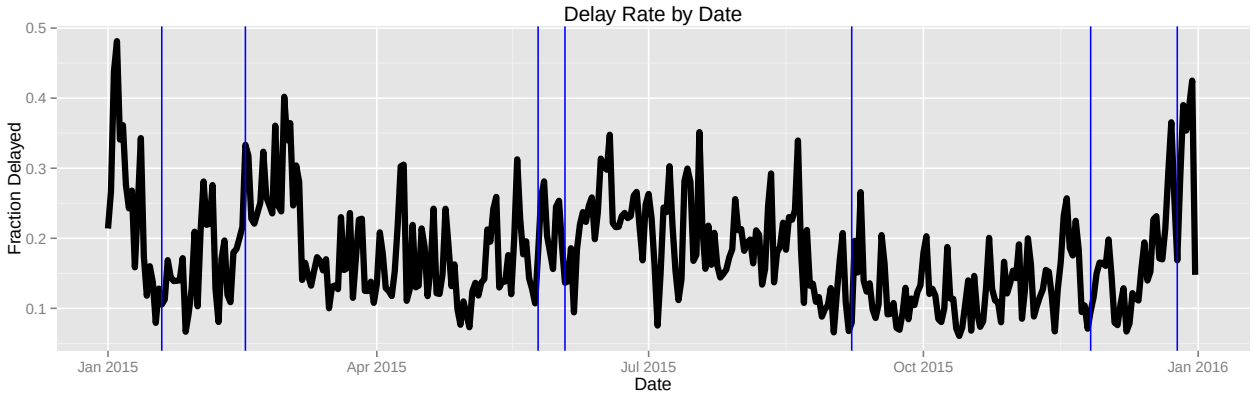
Carrier	Delay Rate
Spirit Air Lines	0.29
Frontier Airlines Inc.	0.25
JetBlue Airways	0.22
American Eagle Airlines Inc.	0.21
United Air Lines Inc.	0.20
Atlantic Southeast Airlines	0.19

Carrier	Delay Rate
Virgin America	0.19
Southwest Airlines Co.	0.18
Skywest Airlines Inc.	0.18
US Airways Inc.	0.18
American Airlines Inc.	0.18
Delta Air Lines Inc.	0.13
Alaska Airlines Inc.	0.12
Hawaiian Airlines Inc.	0.11

Table 8: Delayed by Airport

Airport	Departing Delay Rate	Arriving Delay Rate	Avg Delay Rate
St Cloud Regional	0.31	0.39	0.35
New Castle County	0.39	0.31	0.35
Gustavus	0.43	0.24	0.34
North Bend Muni	0.28	0.31	0.30
Aspen-Pitkin Co/Sardy	0.29	0.30	0.29
Adak	0.41	0.12	0.27
Southeast Texas Regional	0.24	0.27	0.26
Gunnison County	0.23	0.27	0.25
Mammoth Yosemite	0.29	0.21	0.25
Trenton-Mercer County	0.22	0.27	0.24

The plot below shows rate of delay by day, and suggests a spike in delays around Christmas and New Years, as well as an increase in February around and after President's Day. Again, this conflates different causes of delay, which might be interesting to further investigate.



Delay Rate Model

I was interested in building a model to predict the probability that a given flight would be delayed. To do this, I consider each flight in the 2015 data set as a data point, split this set randomly into a training set and holdout set, and trained a regularized logistic regression on the binary dependent variable (delayed or not delayed). As independent variables, I used the airline, day of week, month, and average delay rate for the destination and origin airports in the training set.

The coefficients of the logistic regression model are listed below (Positive coefficients increase the predicted probability of delay, and negative coefficients decrease the predicted probability). Airports with high delay rates as origins or departures are strong positive predictors of delay, and different periods of time and airlines have reasonable coefficients given the charts above.

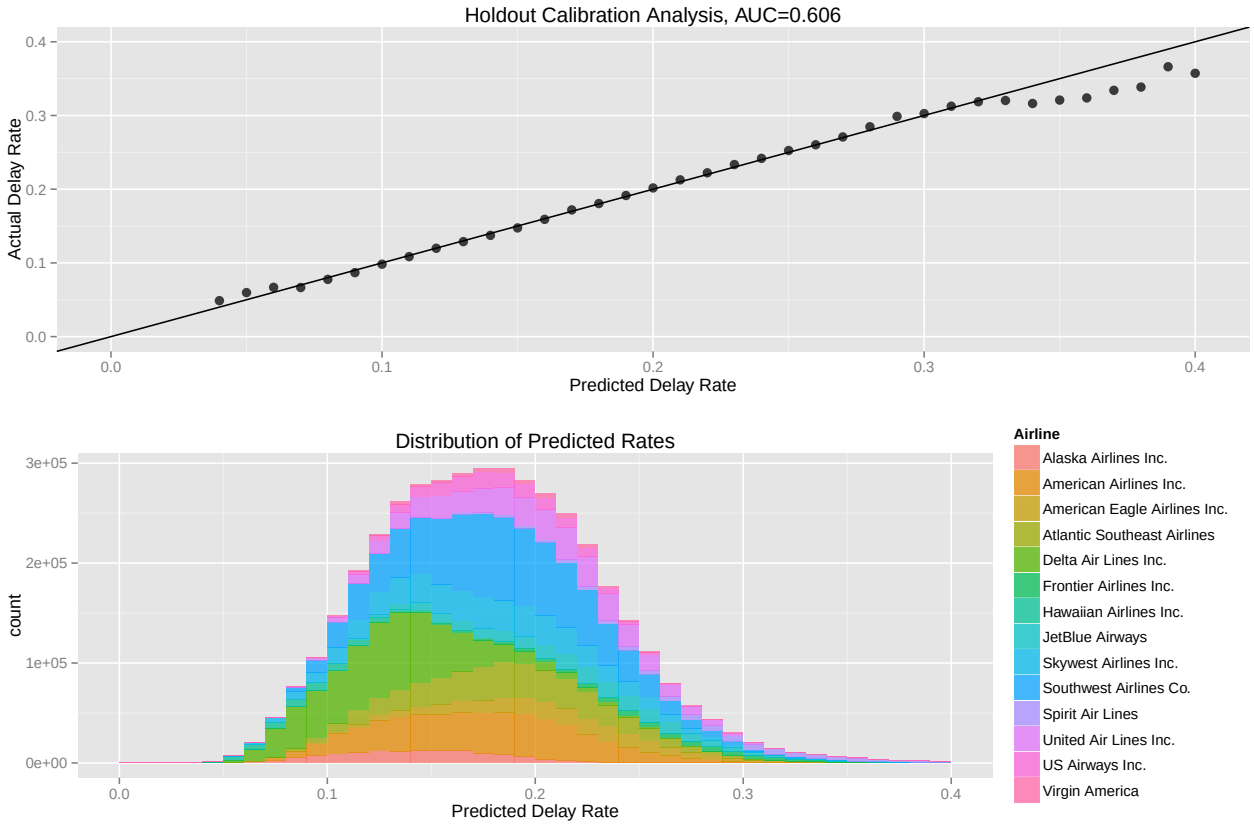
Table 9: Model Coefficients

Feature Name	Coefficients
Origin_delay_rate	5.612
Dest_delay_rate	5.067
(Intercept)	-3.339
month_wordOctober	-0.595
AirlineSpirit Air Lines	0.573
month_wordSeptember	-0.536
AirlineFrontier Airlines Inc.	0.437
month_wordNovember	-0.354
dowSaturday	-0.223
month_wordApril	-0.202
AirlineDelta Air Lines Inc.	-0.196
month_wordJune	0.179
month_wordFebruary	0.167
AirlineJetBlue Airways	0.164
month_wordMay	-0.106
AirlineSkywest Airlines Inc.	0.106
AirlineAmerican Eagle Airlines Inc.	0.104
AirlineSouthwest Airlines Co.	0.087
month_wordAugust	-0.085
AirlineAtlantic Southeast Airlines	0.078
AirlineVirgin America	-0.078
dowTuesday	-0.061
AirlineUS Airways Inc.	-0.057
dowWednesday	-0.055
AirlineUnited Air Lines Inc.	0.052
month_wordMarch	-0.049
dowSunday	-0.048
AirlineAmerican Airlines Inc.	-0.034
dowThursday	0.027
month_wordJuly	0.011
month_wordDecember	0.009
dowFriday	0.000
AirlineHawaiian Airlines Inc.	0.000

The output of the model is summarized in the two plots below. The top plot describes the model performance. This toy model was able to predict delays in the holdout set better than random- the AUC=0.6 describes the ability of the model to rank-order the flights, and the calibration plot shows quantitative agreement between

predicted delay rates and the actual delay rates in the holdout set. The lower plot is the distribution of lateness probability for each flight in the holdout set, colored by airline.

As a next step, a model could be investigated with additional features (simplifying month to season, sorting airports by size, describing the route frequency or distance, proximity to holidays) or using another methodology which might be able to pick up on more subtle interactions between features.



Geographic Visualizations

Below are two plots which provide a visualization of this data. The first plots each airport, with the size proportional to the number of departing flights, and the color indicating the delay rate.

After exploring the data, one point which surprised me is how much air travel is dominated by a small number of airports. Though there are 3245 airports in this data set, the top 5 airports are either the origin or destination for 55% of flights! The second plot illustrates this, with line segments connecting airports for each flight to or from the top 5 airports.